Linear Regression with Gaussian Processes

Manel Martínez-Ramón

ECE, UNM

October, 2018

- Assume a linear estimator

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \quad y = f(\mathbf{x}) + \varepsilon$$

- were ε is the estimation error, and $y \in \mathbb{R}$ are the regressors. The bias is included in the input, which has the form

$$\mathbf{x} = \{1, x_1, \cdots x_d\}^\top$$

The error ε is assumed to be an i.i.d. Gaussian process with zero mean and variance σ_n^2 or, in other words, additive white Gaussian noise (AWGN).

Likelihood

- Now let us take care of the noise process ε and take $f(\mathbf{x})$ as a constant term. Then, y is a Gaussian process with a mean equal to $f(\mathbf{x})$ and a variance σ_n^2.
- The likelihood of sample $y[n]$ given the input $\mathbf{x}[n]$ and the parameters $\mathbf{w}$ is

$$p(y[n]|\mathbf{x}[n], \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{|y[n] - \mathbf{x}[n]^\top \mathbf{w}|^2}{2\sigma_n^2}\right)$$

- We can compute the distribution of the joint process $\mathbf{y}$ by applying the independence assumption. Indeed

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y[n]|\mathbf{x}[n], \mathbf{w})$$

Likelihood

- Then, the likelihood is a joint Gaussian of the form

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{|y[n]-\mathbf{x}[n]^\top \mathbf{w}|^2}{2\sigma_n^2}\right) \\
&= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{|\mathbf{y}-\mathbf{X}^\top \mathbf{w}|^2}{2\sigma_n^2}\right)
\end{aligned}
$$

- Assume now that parameters $\mathbf{w}$ are a linear combination of a data set. In that case, these parameters are also a random process that depends on $\mathbf{X}$ and $\mathbf{y}$.
- We assume that the process $\mathbf{w}$ satisfies the conditions of the Central Limit Theorem: it is a Gaussian random variable, for which the mean is zero.

- We can assume that the prior distribution of the parameters is

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) = \frac{1}{(2\pi|\Sigma_p|)^{(D+1)/2}} \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1} \mathbf{w}\right)$$

where Σ_p is the covariance of the process. It can be shown that this covariance can be arbitrarily set as an identity matrix.

- The posterior with respect to $\mathbf{X}$ and $\mathbf{y}$ is then

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

where the numerator contains the prior and the likelihood, and the denominator, the marginal likelihood.

- Actually, what we need here is to maximize the posterior, this is, to find the set of parameters $\mathbf{w}$ with maximum probability given $\mathbf{X}$ and $\mathbf{y}$, so the denominator is irrelevant because it does not depend on the parameters. Then we can use

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

Hence

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto \exp\left(-\frac{|\mathbf{y}-\mathbf{X}^\top\mathbf{w}|^2}{2\sigma_n^2}\right)\exp\left(-\frac{1}{2}\mathbf{w}^\top\Sigma_p^{-1}\mathbf{w}\right)$$

which is a product of two Gaussians, so it must be a Gaussian.

Prior and posterior on the parameters

- Ignoring the term 1/2 the exponent can be arranged as follows

$$\begin{aligned}
&\sigma_n^{-2}\left(\mathbf{y}-\mathbf{X}^\top\mathbf{w}\right)^\top\left(\mathbf{y}-\mathbf{X}^\top\mathbf{w}\right)-\mathbf{w}^\top\Sigma_p^{-1}\mathbf{w}\\
&=\sigma_n^{-2}\mathbf{y}^\top\mathbf{y}+\mathbf{w}^\top\left(\sigma_n^{-2}\mathbf{X}\mathbf{X}^\top+\Sigma_p^{-1}\right)\mathbf{w}-2\sigma_n^{-2}\mathbf{y}\mathbf{X}^\top\mathbf{w}\\
&=\sigma_n^{-2}\mathbf{y}^\top\mathbf{y}+\mathbf{w}^\top\mathbf{A}\mathbf{w}-2\sigma_n^{-2}\mathbf{y}\mathbf{X}^\top\mathbf{w}
\end{aligned}$$

with $\mathbf{A}=\sigma_n^{-2}\mathbf{X}\mathbf{X}^\top+\Sigma_p^{-1}$. The expression of the Gaussian must have an exponent

$$\frac{1}{2}\left(\mathbf{w}-\bar{\mathbf{w}}\right)^\top\mathbf{A}\left(\mathbf{w}-\bar{\mathbf{w}}\right)$$

where $\bar{\mathbf{w}}$ and $\mathbf{A}^{-1}$ play the role of a mean and a covariance.

- If we equal both expressions and simplify the terms (again ignoring the term 1/2)

$$\begin{aligned}\sigma_n^{-2}\mathbf{y}^\top\mathbf{y} + \mathbf{w}^\top\mathbf{A}\mathbf{w} - 2\sigma_n^{-2}\mathbf{y}\mathbf{X}^\top\mathbf{w} &= (\mathbf{w}-\bar{\mathbf{w}})^\top\mathbf{A}(\mathbf{w}-\bar{\mathbf{w}})\\ \sigma_n^{-2}\mathbf{y}^\top\mathbf{y} - 2\sigma_n^{-2}\mathbf{y}\mathbf{X}^\top\mathbf{w} &= \bar{\mathbf{w}}^\top\mathbf{A}\bar{\mathbf{w}} - 2\bar{\mathbf{w}}^\top\mathbf{A}\mathbf{w}\end{aligned}$$

- Then, necessarily

$$\sigma_n^{-2}\mathbf{y}\mathbf{X}^\top = \bar{\mathbf{w}}^\top\mathbf{A}$$

  and

$$\bar{\mathbf{w}} = \sigma_n^{-2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y}$$

  which, in turn, satisfies $\bar{\mathbf{w}}^\top\mathbf{A}\bar{\mathbf{w}} = \sigma_n^{-2}\mathbf{y}^\top\mathbf{y}$

Prior and posterior on the parameters

- Finally, multiplying again by 1/2

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{w}-\bar{\mathbf{w}}\right)^{\top}\mathbf{A}\left(\mathbf{w}-\bar{\mathbf{w}}\right)\right)$$

where

$$\mathbf{A} = \sigma_n^{-2}\mathbf{X}\mathbf{X}^{\top} + \Sigma_p^{-1}$$

is the inverse of the covariance, and

$$\bar{\mathbf{w}} = \left(\mathbf{X}\mathbf{X}^{\top} + \sigma_n^2\Sigma_p^{-1}\right)^{-1}\mathbf{X}\mathbf{y}$$

- This result is exactly equal to the ridge regression if $\Sigma_p^{-1} = \mathbf{I}$
- The optimal value for σ_n^2 can be estimated by Maximum Likelihood, as we will see in next lessons.

Predictive likelihood

- Assume that a new sample $\mathbf{x}^*$, not belonging to the training set $\mathbf{X}$, is available. The estimator will produce a prediction

$$f_* = \mathbf{w}^\top \mathbf{x}^*$$

Using the expression of slide 4, we can compute the likelihood of f_* given the new sample $\mathbf{x}^*$ and a particular value of $\mathbf{w}$, which can be expressed as

$$p(f_*|\mathbf{x}, \mathbf{w})$$

We also have the posterior on $\mathbf{w}$. Using the Total Probability Theorem we have

$$p(f_*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int_{\mathbf{w}} p(f_*|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{y}, \mathbf{X}) d\mathbf{w}$$

Predictive likelihood

- Solving the integral we have

$$p(f_*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\bar{\mathbf{w}}^\top \mathbf{x}^*, \mathbf{x}^{*\top} \mathbf{A}^{-1} \mathbf{x}^*\right)$$

- The advantage of the Gaussian Process over the standard MMSE or Ridge Regression is that now we have a distribution on the prediction. In other words, we can judge how accurate is our prediction just taking a look to the variance $\sigma_{f_*}^2 = \mathbf{x}^{*\top} \mathbf{A}^{-1} \mathbf{x}^*$ of the output.
- Finally, the whole method can be kernelized and we can still make inference and obtain a predictive likelihood under the Gaussian hypothesis.

Outcomes of this lesson

In this lesson we have introduced the liner Gaussian Process for regression. The main aspects to retain are:

- The concept of regression.
- The idea of data likelihood: the probabilistic model for y_n.
  - We assume that y_n is iid: joint likelihood as product of likelihoods.
- $\mathbf{w}$ is treated as a latent random variable with a Gaussian prior.
- The posterior is proportional to this prior times the likelihood (Bayes rule).
- WIth the posterior and the Total Probablity rule, we find the posterior of the predictions.